Classification Tree

- In regression, the predicted response for an observation is given by mean response of training observations that belongs to same terminal node.
- In classification, we can predict that each observation belong to most commonly occuring class of training observations in the region which it belongs.
- While splitting node, we check classification error rate. Classification error rate is fraction of the training observation in that region that do not belong to the most common class.
- 2 measures are used for this – i) Gini Index ii) Cross entropy.
- Gini Index,

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk})$$

- A total variance about the K classes. Gini Index takes on small values if all the $\hat{p}_{mk}$ are close to one or zero.
- For this reason Gini Index are referred to a measure of node purity - a small value indicates that a node contains predominantly observation to a single class.
- Cross entropy,

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Cross entropy will take a value zero or near to zero if $\hat{p}_{mk}$ are all near to zero or one.

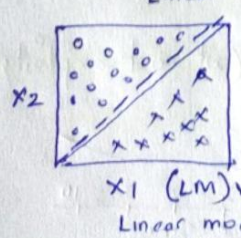
- So if the node is pure or homogenous, both Cross entropy and Gini Index take the value near to zero.

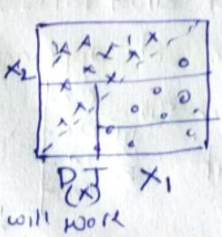
TREES VS LINEAR MODEL (WHICH ONE IS BETTER)

- If the relationship between the features and the response is well approximate by linear model then linear regression will work well.
- If there is non-linear highly and complex relationship between feature and response variable then decision tree will work well.
- Interpretation is easy in case of Tree based model.

Consider one Independent and one dependent variable X_1 and X_2.

Linear Relationship





Linear model will work

Non Linear Relationship